A Generalized Fuzzy Linguistic Model for Predicting Component Concentrations in an Optical Gas Sensing System

Wang, Y., Cao, H., Yan, X., Zhou, Y., Liu, X., & McLoone, S. (2016). A Generalized Fuzzy Linguistic Model for Predicting Component Concentrations in an Optical Gas Sensing System. Chemometrics and Intelligent Laboratory Systems, 158, 21-30. DOI: 10.1016/j.chemolab.2016.07.012

Published in:
Chemometrics and Intelligent Laboratory Systems

Document Version:
Peer reviewed version

A Generalized Fuzzy Linguistic Model for Predicting Component Concentrations in an Optical Gas Sensing System

Yanxia Wang[a,c], Hui Cao[a], Xingyu Yan[a], Yan Zhou[b,*], Xueqin Liu[c], Seán McLoone[c]

[a]*State Key Laboratory of Electrical Insulation and Power Equipment, School of Electrical Engineering, Xi'an Jiaotong University, Xi'an 710049, China*
[b]*School of Energy and Power Engineering, Xi'an Jiaotong University, Xi'an 710049, China*
[c]*School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast BT9 5AH, U.K.*

Abstract

Motivated by environmental protection concerns, monitoring the flue gas of thermal power plant is now often mandatory due to the need to ensure that emission levels stay within safe limits. Optical based gas sensing systems are increasingly employed for this purpose, with regression techniques used to relate gas optical absorption spectra to the concentrations of specific gas components of interest (NO_x, SO_2 etc.). Accurately predicting gas concentrations from absorption spectra remains a challenging problem due to the presence of nonlinearities in the relationships and the high-dimensional and correlated nature of the spectral data. This article proposes a generalized fuzzy linguistic model (GFLM) to address this challenge. The GFLM is made up of a series of "If-Then" fuzzy rules. The absorption spectra are input variables in the rule antecedent. The rule consequent is a general nonlinear polynomial function of the absorption spectra. Model parameters are estimated using least squares and gradient descent optimization algorithms. The performance of GFLM is compared with other traditional prediction models, such as partial least squares, support vector machines, multilayer perceptron neural networks and radial basis function neural networks, for two real flue gas spectral datasets: one from a coal-fired power plant and one from a gas-fired power plant. The experimental results show that the generalized fuzzy linguistic model has good predictive ability, and is competitive with alternative approaches, while having the added advantage of providing an interpretable model.

Keywords: Optical Gas Sensing Systems, Generalized Fuzzy Linguistic Model, Parameter Optimization.

1. Introduction

In order to demonstrate compliance with regularity requirements on thermal power plant emissions, monitoring the concentrations of pollutants such as nitrogen oxides, sulphur oxides and oxycarbides in flue gas emissions is now mandatory in many countries [1][2]. Gas

*corresponding author: yan.zhou@mail.xjtu.edu.cn

sensing methods are diverse due to the chemical and physical effects that can reflect gas characteristics [3]. One common sensing principle is the electrochemical variations that occur between target gases and different sensor materials, such as metal oxide semiconductors, polymers, and carbon nanotubes [4][5]. In recent decades, optical spectroscopy based methods have become increasingly popular for gas sensing [6][7][8][9][10], due to their high sensitivity, selectivity and stability. These methods measure the chemical composition dependent absorption of light that occurs at different wavelengths when light passes through the flue gas [11]. By analyzing the measured absorption spectra, the concentration of specific components of the gas can be predicted by regression models [12].

Many regression methods for spectral data have been reported [13]. As a well-known multivariate regression algorithm, the classical partial least squares (PLS) can only establish linear relationships between absorption spectra and component concentrations [14][15][16]. In experiments, however, there are many conditions that can lead to nonlinearity such as instrument variation and analyte characteristics [17]. Nonlinear modelling methods such as multilayer perceptron (MLP) neural networks [18], radial basis function (RBF) networks [19], and support vector machines (SVM) [20][21] can be used to learn the nonlinear relationships. However, these methods typically require substantial computational effort to train, and by virtue of their black-box structure, cannot provide understandable heuristic knowledge [22].

Linguistic models are built up by fuzzy rules that express human-readable descriptions in a format suitable for regression analysis [23][24]. A fuzzy rule is a logical linguistic "If-Then" statement [25], where the "If" expression is referred to as the antecedent and the "Then" expression as the consequent. The antecedent expresses input conditions in terms of fuzzy linguistic labels. Two forms of consequent are normally employed in fuzzy models; the first expresses the output directly as linguistic labels and is referred to as the Mamdani fuzzy rule [26], while the second defines the output as a linear function of the inputs and is called the Takagi-Sugeno formulation. The latter is preferred for modelling applications because it produces a crisp output without defuzzification, and yields reduced complexity regression models [27][28]. In our previous work, we discussed a technology with a series of Takagi-Sugeno fuzzy rules for quantitative analysis [29]. Considering the nonlinearity of spectral data, we proposed a quadratic polynomial equation as the rule consequent [30]. Nevertheless, the predefined form of the rule consequent employed may limit the approach's power to handle variation in nonlinear complexity for different chemical concentration estimation tasks.

In this article, a generalized fuzzy linguistic model (GFLM) suited to the optical gas sensing system modeling problem is presented. The model consists of a sequence of If-Then fuzzy rules. In the rule antecedent, the input variables are absorption spectra. The rule consequent is a general nonlinear polynomial function expressed as a function of the absorption spectra. Least squares and gradient descent are both adopted to optimize the model. To demonstrate the performance of GFLM, it is compared with PLS, SVM, MLP and RBF models for flue gas spectral datasets from coal-fired and gas-fired power plants.

The reminder of the paper is organized as follows. The optical gas sensing system and GFLM are described in section 2. In section 3, the experimental setup (datasets and procedure) is described in detail, while section 4 presents and discusses the experimental

results. Finally, section 5 concludes the paper.

2. Gas Sensing System with a Generalized Fuzzy Linguistic Model

2.1. Optical Gas Sensing System

The schematic diagram of an optical gas sensing system is shown in Figure 1. Flue gas is drawn into an explosion-proof tubular heater and heated to a predefined temperature. It is then transferred to an absorption cell where lights from a known light source is shone through the gas onto miniature spectrometers which measure the absorption spectrum.

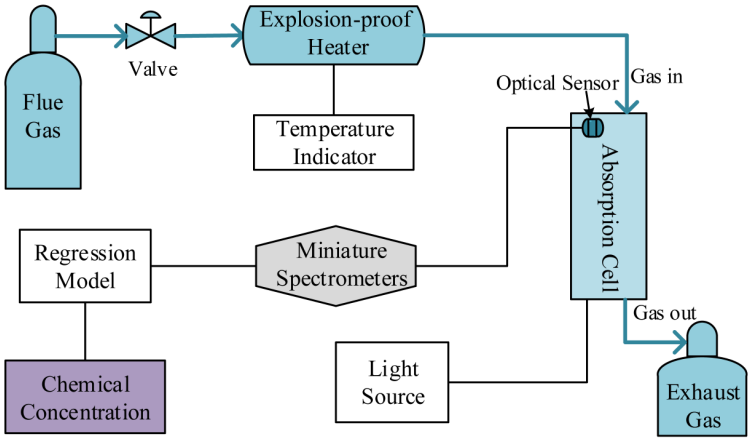

Figure 1: Optical gas sensing system

2.2. Generalized Fuzzy Linguistic Model (GFLM)

GFLM is functionally equivalent to a series of logical "If-Then" fuzzy rules. The antecedent "If" presents the conditions, using fuzzy linguistic labels instead of crisp numbers. The consequent "Then" is a general nonlinear polynomial function expressed in terms of the input variables. The "If-Then" fuzzy rules thus assume the form:

If u_1 is A_1 and u_2 is A_2 and $\cdots$ and u_n is A_n

$$\text{Then } f_i = \sum_{j_1=1}^{n} k_{j_1}^1 u_{j_1} + \sum_{j_1=1}^{n}\sum_{j_2=j_1}^{n} k_{j_1 j_2}^2 u_{j_1} u_{j_2} + \cdots + \sum_{j_1=1}^{n}\sum_{j_2=j_1}^{n}\cdots\sum_{j_m=j_{m-1}}^{n} k_{j_1 j_2 \cdots j_m}^m u_{j_1} u_{j_2} \cdots u_{j_m} + b$$

where $\{u_1, u_2, ...u_n\}$ is the input vector, $\{A_1, A_2, ...A_n\}$ are the linguistic labels, $\{k_{j_1}^1, k_{j_1 j_2}^2, ..., k_{j_1 j_2 \cdots j_m}^m, b\}$ is the vector of consequent parameters, and m is the highest degree considered. When $m = 1$, this fuzzy rule reduces to the classical Takagi-Sugeno type rule. A given GFLM consists of a series of these rules, each one having the same highest polynomial degree, m.

To initialize the model, a clustering technique is first used to determine the initial locations of the linguistic labels. The number of clusters R can be predefined by users or automatically determined as part of the clustering process.

Gaussian functions are employed to generate the membership degree of each linguistic fuzzy set. The firing strength of the i-th rule is then calculated as:

$$\omega_i = \prod_{j=1}^{n} \mu(u_j) = \prod_{j=1}^{n} e^{-\frac{(u_j - c_{i,j})^2}{2\sigma_{i,j}^2}} \tag{1}$$

where $\prod$ performs fuzzy AND, and $\{c_{i,j}, \sigma_{i,j}\}$ is the antecedent parameter set. The Gaussian function varies when these parameters change, thus exhibiting various firing strengths. The output is computed as:

$$y = \frac{\sum_{i=1}^{R} \omega_i f_i}{\sum_{i=1}^{R} \omega_i} = \sum_{i=1}^{R} \bar{\omega}_i f_i \tag{2}$$

where f_i is the output of i-th rule.

Thus, we have constructed the generalized fuzzy linguistic model. Next, a learning procedure needs to be developed. For simplicity, we assume that the parameters can be decomposed into a nonlinear set $S_N = \{c_{i,j}, \sigma_{i,j}\}$ and linear set $S_L = \left\{k_{j_1}^1, k_{j_1 j_2}^2, ..., k_{j_1 j_2 \cdots j_m}^m, b\right\}$. Now given values of the elements of S_N, we can determine estimates for S_L by solving:

$$XS_L = Y \tag{3}$$

where X is a regressor matrix whose elements are a function of S_N and the model inputs, i.e. $X = [x_{kj}] = g_j(S_N, u_1(k), u_2(k), \cdots, u_n(k))$. The linear least squares solutions to (3), which minimizes $||XS_L - Y||^2$, is given by:

$$\hat{S}_L = (X^T X)^{-1} X^T Y \tag{4}$$

While (4) is concise in notation, $X^T X$ can often be ill-conditioned or singular leading to numerical issues if computed directly; singular value decomposition (SVD) provides a stable approach to address this [31]. However, determining the solution in this manner is computationally expensive which can be an issue if X is large. Alternatively, S_L can be computed using the recursive least squares estimator [32][33], defined as:

$$\begin{cases} S_L(n+1) = S_L(n) + P_{n+1} x_{n+1} (y_{n+1} - x_{n+1}^T S_L(n)) \\ P_{n+1} = P_n - \dfrac{P_n x_{n+1} x_{n+1}^T P_n}{1 + x_{n+1}^T P_n x_{n+1}} \end{cases} \tag{5}$$

where n ranges from 1 to N, N is the number of training data pairs and the desired $\hat{S}_L$ is given by $S_L(N)$. The initial conditions needed are $S_L(1) = 0$ and $P_0 = \lambda I$, where λ is a positive large number and I is the identity matrix.

Once the linear elements in S_L are computed, gradient descent can be used to update the nonlinear elements in S_N in order to minimize an error measure defined over the training

data [34], that is

$$E = \sum\nolimits_{p=1}^{N} E_p = \sum\nolimits_{p=1}^{N} \left(y_p^d - y_p\right)^2 \tag{6}$$

where y_p^d is the desired output and y_p is the actual output of the GFLM.

The derivative of the error measure is used as the gradient vector. For $\alpha \in S_N$, the derivative of the overall error measure E with respect to α is

$$\frac{\partial E}{\partial \alpha} = \sum_{p=1}^{N} \frac{\partial E_p}{\partial \alpha} = -2 \sum_{p=1}^{N} \left(y_p^d - y_p\right) \cdot \frac{\partial y_p}{\partial \alpha} \tag{7}$$

Accordingly, the update formula for the parameter α is

$$\Delta\alpha = -\eta \frac{\partial E}{\partial \alpha} \tag{8}$$

where η is the learning rate, which can be further written as

$$\eta = \frac{\kappa}{\sqrt{\sum\limits_{\alpha \in S_1} \left(\frac{\partial E}{\partial \alpha}\right)^2}} \tag{9}$$

where κ is the step size, the length of each transition along the negative gradient direction in the parameter space.

For $\alpha = c_{ij}$, the derivative $\frac{\partial y_p}{\partial \alpha}$ in equation (7) is given by:

$$\begin{aligned} \frac{\partial y_p}{\partial c_{ij}} &= \frac{\omega_i (f_i \sum_{i=1}^{R} \omega_i - \sum_{i=1}^{R} \omega_i f_i)}{\left(\sum_{i=1}^{R} \omega_i\right)^2} \cdot \frac{(x_{ij} - c_{i,j})}{\sigma_{i,j}^2} \\ &= \frac{\omega_i \sum\limits_{\substack{k=1 \\ k \neq i}}^{R} (f_i - f_k)\omega_k}{\left(\sum_{i=1}^{R} \omega_i\right)^2} \cdot \frac{(x_{ij} - c_{i,j})}{\sigma_{i,j}^2} \end{aligned} \tag{10}$$

and for $\alpha = \sigma_{ij}$, it is defined as:

$$\begin{aligned} \frac{\partial y_p}{\partial \sigma_{ij}} &= \frac{\omega_i (f_i \sum_{i=1}^{R} \omega_i - \omega_i f_i)}{\left(\sum_{i=1}^{R} \omega_i\right)^2} \cdot \frac{(x_{ij} - c_{i,j})^2}{\sigma_{i,j}^3} \\ &= \frac{\omega_i \sum\limits_{\substack{k=1 \\ k \neq i}}^{R} (f_i - f_k)\omega_k}{\left(\sum_{i=1}^{R} \omega_i\right)^2} \cdot \frac{(x_{ij} - c_{i,j})^2}{\sigma_{i,j}^3} \end{aligned} \tag{11}$$

The least squares estimate and gradient descent are performed iteratively until the num-

ber of iterations reaches a predefined value or the error E is less than a specified threshold.

3. Experimental Setup

3.1. Datasets

Case study I:Coal-fired power plant flue gas

Spectral data for the flue gas from a coal-fired power plant was collected using miniature spectrometers (USB2000+XR1) from Ocean Optics. In total 196 samples were measured and for these samples, the concentration ranges of nitric oxide, sulfur dioxide and nitrogen dioxide, measured using a Testo 350 industrial flue gas analyzer, varied in the range 0-2000, 0-2500 and 0-500ppm, respectively. Using the ocean optics UV-VIS-NIR light source DH-2000, the wavelength range is from 187.87nm to 1026.97nm with a resolution of 0.82 nm; thus the dimension of each dataset is 1024. The absorption spectra are shown in Figure 2.

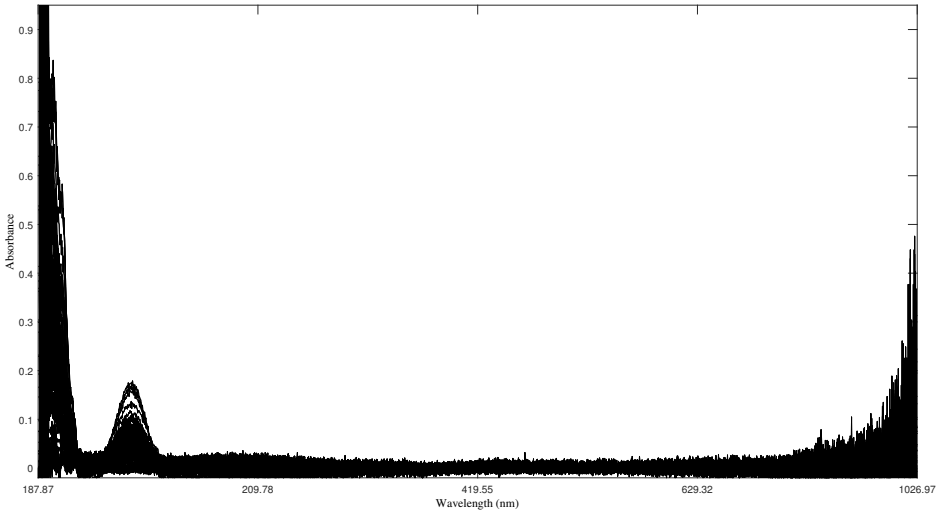

Figure 2: The absorption spectra of flue gas from a coal-fired power plant

Case study II:Gas-fired power plant flue gas

This dataset, plotted in Figure 3, consisted of spectral data for 198 samples of flue gas collected from a gas-fired power plant using a GASMET DX4000 Fourier transform infrared gas analyzer, which measures light intensity with the wavenumber ranging from $734.65cm^{-1}$ to $4191.98cm^{-1}$ with a resolution of $7.72cm^{-1}$; thus the dimension of each dataset is 448. For each sample, the chemical concentrations were measured using a SP-3400 gas chromatograph from Beijing Beifen-Ruili Analytical Instrument Co. Ltd. Concentrations of a mixture of methane, carbon monoxide and carbon dioxide were recorded and the observed ranges over the 198 samples were 0-5100, 0-4500, and 0-6000 ppm, respectively.

Table 1 provides a detailed description of these two datasets.

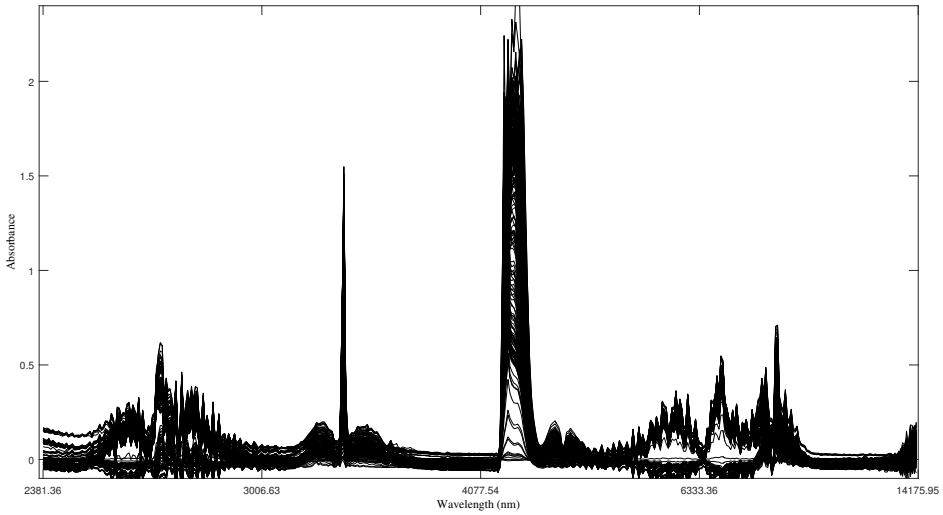

Figure 3: The absorption spectra of flue gas from a gas-fired power plant

Table 1: The Flue Gas Datasets

Dataset	Component	Calibration			Test		
		No.	Conc. Range (ppm)	Mean Conc. (ppm)	No.	Conc. Range (ppm)	Mean Conc. (ppm)
Coal-fired Flue Gas	Nitric Oxide	132	0-2000	875.59	66	0-2000	828.50
	Sulfur Dioxide		0-2500	528.55		0-2500	605.31
	Nitrogen Dioxide		0-500	98.20		0-500	111.72
Gas-fired Flue Gas	Methane	131	0-5100	2046.37	65	0-5100	2209.13
	Carbon Monoxide		0-4500	981.16		0-4500	917.74
	Carbon Dioxide		0-6000	1371.64		0-6000	1319.47

3.2. Experimental Procedure

To evaluate the performance of the proposed GFLM model structure, GFLMs with rule consequent polynomial degrees ranging from 1 to 5 were estimated and compared with PLS, SVM, MLP and RBF models for the task of predicting gas concentrations from the recorded flue gas optical spectra. In our experiments, the datasets were split into calibration and test sets with every third sample chosen for testing and the remaining samples used for the calibration set [35]. The calibration sets were used to train the regression models, and the test sets used to evaluate their effectiveness.

The PLS, SVM, MLP and RBF methods were developed using the implementations in MATLAB version 7.11.0 (R2013b) with default parameters employed, unless otherwise specified. PLS adopts a multi-input single-output (MISO) model structure. For SVM models, the Gaussian function was used as the kernel function. Three layer topologies were used with

both the MLP and RBF networks with sigmoid functions used as the activation function in the hidden layer and a linear function employed in the output layer in the MLP network and radial basis functions used as the nonlinear activation function in the RBF network. Using the cross-validation root mean square error (CV-RMSE) as a performance metric, 10-fold cross-validation was employed to optimize the hyperparameters of each modelling paradigm [36]. For PLS, the maximum number of latent variables (LVs) considered was 30. For SVM models, the penalty parameter (C) was selected in the range [-4, 15] in a log 2 space (the normal space ranging from 2^{-4} to 2^{15}), and the kernel parameter ($1/\sigma^2$) of the Gaussian was determined using a grid search in the range [-5, 6] in log 2 space (the normal space ranging from 2^{-5} to 2^{6}). The optimal number of hidden nodes in the MLP and RBF were chosen from the range [1, 10] and [1, 30] respectively. To reduce the computational complexity of regression models, and the dimensionality of the input space relative to the number of samples, the principal components of the absorption spectra were extracted and used as input variables to the model with the number of components retained capturing 90% of the observed variance [37][38]. This was found to be 5 components for both the coal-fired power plant and gas-fired power plant flue gas data sets. The clustering algorithm used to initialize the GFLM model was subtractive clustering [39]. In this algorithm, the tuning parameter 'radii', which determines the number of clusters generated was varied between 0.1 and 0.9 in steps of 0.1.

In this article, the CV-RMSE, the prediction root mean squared error (P-RMSE), the relative error of prediction expressed as a percentage (REP%) and the randomization t-test [40] are employed as metrics to compare and assess the efficacy of various regression models. The randomization t-test is applied to detect statistically significant differences in model performance. The significance level is set as 0.10 in accordance with [41]. The cross-validation metrics are computed with respect to the calibration datatsets, while the prediction metrics are computed with respect to the test datasets. All the regression models are implemented in MATLAB 7.11.0 on a general-personal computer with an Intel i7-2600 CPU and 8 GB of RAM.

4. Results and Discussion

4.1. The coal-fired power plant flue gas dataset

The hyperparameter cross-validation errors obtained with PLS, MLP, RBF, SVM and GFLM models for nitric oxide, sulfur dioxide and nitrogen dioxide are illustrated in Figures 4, 5 and 6, respectively. The hyperparmeters that yielded the minimum CV-RMSE were selected as the optimum values for each model. The analytical results are summarized in Table 2. For nitric oxide, the P-RMSE value obtained with GFLM is 2.22%, 36.29%, 4.43% and 8.92% lower than those of the PLS, SVM, MLP and RBF, respectively. The REP% value of GFLM is 13.68%, 36.22%, 11.74% and 30.25% lower than those of PLS, SVM, MLP and RBF respectively. Application of the randomization t-test confirms that the better predictive ability of GFLM is statistically significant for all model ($p < 0.1$).

For sulfur dioxide, while the REP% for GFLM is marginally inferior to PLS, it is superior in terms of the other metrics. The P-RMSE with GFLM is 9.1%, 282.62%, 248.14% and

194.37% smaller than those achieved by the PLS, SVM, MLP and RBF models, respectively. Statistically, the difference in performance between PLS and GFLM is not significant ($p = 0.53$), while the differences between the other models and GFLM are very significant ($p < 0.005$).

For nitrogen dioxide, the P-RMSE value of GFLM is 1.29%, 77.36%, 13.17% and 158.39% lower than those of the PLS, SVM, MLP and RBF respectively. The REP% value of GFLM is 1.91%, 37.36%, 25.69% and 51.94% smaller than those of PLS, SVM, MLP and RBF respectively. The t-test indicates that the improvement in performance with GFLM is statistically significant for all models apart from PLS.

In addition, for GFLM, the highest polynomial degree in the rule consequent is selected to be 1, 2 and 2 for nitric oxide, sulfur dioxide and nitrogen dioxide, respectively. The parameter of subtractive clustering, radii, is chosen as 0.2, 0.9 and 0.8 for nitric oxide, sulfur dioxide and nitrogen dioxide, respectively.

Based on this study, while PLS has similar predictive capability to GFLM for sulfur dioxide and nitrogen dioxide, overall GFLM provides the most consistent performance and is therefore recommended for regression modeling for the coal-fired power plant flue gas.

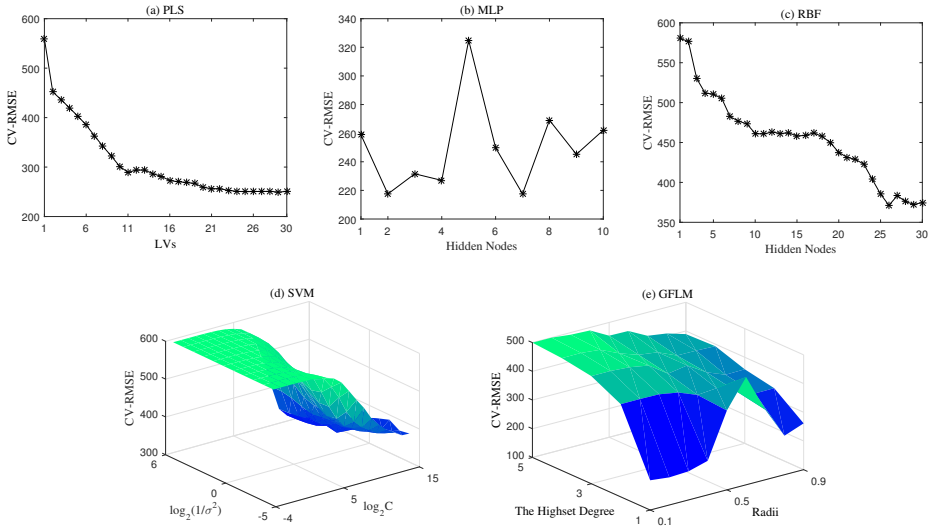

Figure 4: The model hyperparameter cross-validation errors for nitric oxide

4.2. The gas-fired power plant flue gas dataset

The hyperparameter cross-validation errors obtained with PLS, MLP, RBF, SVM and GFLM for methane, carbon monoxide and carbon dioxide are depicted in Figure 7, 8 and 9, respectively. The analytical results for the gas-fired power plant flue gas are summarized in Table 3. For methane, the P-RMSE value with GFLM is 0.67%, 0.34%, 0.3% and 1.4% lower than those of PLS, SVM, MLP and RBF respectively. The REP% of GFLM is 5.56%, 0.8%, 7.25% and 25.92% smaller than those of PLS, SVM, MLP and RBF respectively. Despite

Table 2: Analytical results for the coal-fired power plant flue gas

Component	Model	CV-RMSE	P-RMSE	REP%	Rand.-t	Other parameters
Nitric Oxide	PLS	249.44	235.49	22.85%	0.075	29^a
	SVM	315.65	313.98	27.38%	0.005	$(2^{11})^b$; $(2^6)^c$
	MLP	217.52	240.57	22.46%	0.045	2^d
	RBF	256.90	250.94	26.18%	0.015	26^e
	GFLM	196.21	230.38	20.10%	\	1^f;0.2^g
Sulfur Dioxide	PLS	72.19	27.47	3.07%	0.53	26^a
	SVM	78.48	96.35	13.51%	0.005	$(2^{12})^b$; $(2^{-3})^c$
	MLP	72.86	87.67	11.32%	0.005	6^d
	RBF	87.94	74.12	8.71%	0.005	29^e
	GFLM	70.95	25.18	3.11%	\	2^f;0.9^g
Nitrogen Dioxide	PLS	24.78	22.06	22.37%	0.54	29^a
	SVM	36.71	38.62	30.15%	0.005	$(2^7)^b$; $(2^2)^c$
	MLP	39.93	24.65	27.59%	0.04	7^d
	RBF	53.30	56.27	33.35%	0.005	28^e
	GFLM	26.27	21.78	21.95%	\	2^f;0.8^g

[a] the number of PLS latent variables.
[b] the penalty parameter in SVM.
[c] the kernel parameter of the Gaussian kernel in SVM.
[d] the number of hidden layer nodes in the MLP model.
[e] the number of hidden layer nodes in the RBF model.
[f] the highest degree of polynomial in the GFLM.
[g] the radii parameter of the subtractive clustering algorithm used in the GFLM.

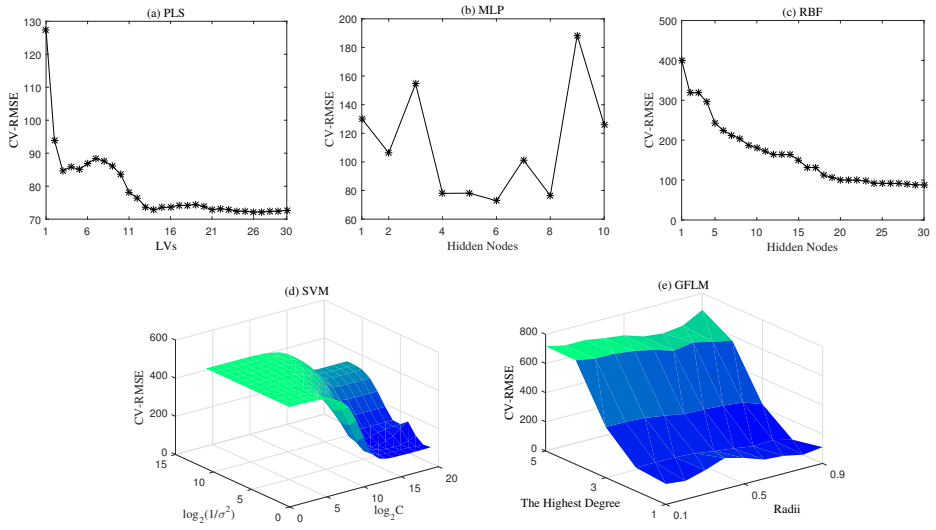

Figure 5: The model hyperparameter cross-validation errors for sulfur dioxide

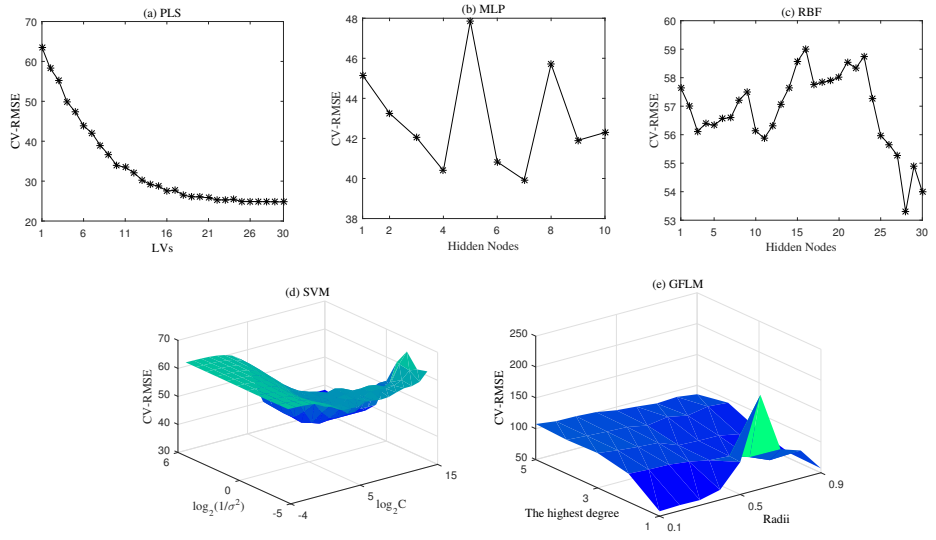

Figure 6: The model hyperparameter cross-validation errors for nitrogen dioxide

the differences in REP% between PLS, SVM, MLP and GFLM, the randomization t-test shows that they are not statistically significant. This is consistent with the observation that the P-RMSE values are very similar for all models in this instance.

For carbon monoxide, PLS offers the best predictive capability in terms of CV-RMSE, whereas GFLM has the smallest P-RMSE value, 42.58%, 375.58%, 112.18% and 117.63% smaller than those achieved by the PLS, SVM, MLP and RBF models respectively. In addition, the REP% of GFLM is 9.44%, 16.67%, 4.46% and 19.02% lower than those of PLS,

SVM, MLP and RBF, respectively. All are statistically significant differences ($p = 0.005$).

For carbon dioxide, the P-RMSE value of GFLM is 1.26%, 46.5%, 21.45% and 33.63% lower than those of the PLS, SVM, MLP and RBF models respectively. The REP% value of GFLM (REP=3.73%) is the lowest, 26.54%, 238.87%, 160.05% and 127.08% smaller than those of PLS, SVM, MLP and RBF, respectively. The t-tests confirm that the superior predictive capability of GLFM is statistically significant with respect to all models.

For GFLM, the highest degree of polynomial in the rule consequents is selected to be 1, 1 and 2 for methane, carbon monoxide and carbon dioxide respectively. The subtractive clustering parameter, radii, is chosen as 0.9 for all 3 models.

Overall, the experimental results verify that the GFLM model is the most effective regression model for the gas-fired power plant flue gas.

4.3. General Observations

It is noteworthy that the REP% values are generally high for both datasets across all models. They are greater than 20% for nitric oxide and nitrogen dioxide no matter which strategy is employed. They are also particularly poor for carbon monoxide the best performing model, GFLM, only achieving an REP% of 21%. These high REP% are due to a substantial level of measurement errors and outliers in the data. A number of factors contribute to this linked to limitations of the measurement equipment and the manual measurement process. In particular, the measurement accuracy of the Testo 350 flue gas analyzer is limited to ±5% and the SP-3400 gas chromatograph, which is used to measure the chemical concentrations, has maximal measurement errors of the order of 10%. The GASMET DX4000 Fourier transform infrared gas analyzer has a measurement accuracy of ±3%. In addition, it is susceptible to gas pressure and temperature changes, which affect the line shape of the measured absorbance spectrum, and thus the accuracy of the analysis results [42]. Spectral noise may also make the analysis less precise. Finally, the manual measurement process is likely to have introduced random errors.

While spectral data is inherently nonlinear, in our experiments the nonlinear ANN models are inferior to the linear PLS models for most components. The poor performance of the ANN models may be related to the size of the experimental datasets available (number of training samples) relative to its dimensionality and the level of noise. These circumstances make estimating ANN models with good generalisation capabilities challenging, particularly with flexible structures such as MLPs and RBFs [43], with the result that simpler model structures (such as the PLS model) can outperform them. The characteristics of the GFLM paradigm, which essentially involves the estimation of relatively simple local models and the interpolation functions that interpolate between them, make it a more robust paradigm in these circumstances, hence its good all round performance.

5. Conclusions

This paper proposes a generalized fuzzy linguistic model (GFLM) to predict chemical concentrations from gas optical absorption spectra. For the GFLM model, the rule consequent is a general nonlinear polynomial function of input variables to characterize high-dimensional data. Least squares and gradient descent optimization algorithms are adopted

Table 3: Analytical results for the gas-fired power plant flue gas

Component	Model	CV-RMSE	P-RMSE	REP%	Rand.-t	Other parameters
Methane	PLS	43.18	86.37	10.17%	0.505	26^a
	SVM	80.57	86.09	9.71%	0.8	$(2^{15})^b$; $(2^{-3})^c$
	MLP	74.79	86.05	10.33%	0.84	7^d
	RBF	79.64	87.00	12.13%	0.025	28^e
	GFLM	64.05	85.80	9.63%	\	1^f; 0.9^g
Carbon Monoxide	PLS	114.43	153.87	23.30%	0.005	30^a
	SVM	508.20	513.24	24.84%	0.005	$(2^{11})^b$; $(2^{-2})^c$
	MLP	257.08	228.98	22.24%	0.005	5^d
	RBF	246.08	234.86	25.34%	0.005	5^e
	GFLM	133.55	107.92	21.29%	\	1^f; 0.9^g
Carbon Dioxide	PLS	83.82	74.59	4.72%	0.08	11^a
	SVM	108.28	107.92	12.64%	0.005	$(2^{15})^b$; $(2^{-3})^c$
	MLP	120.71	89.47	9.70%	0.04	5^d
	RBF	133.14	98.45	8.47%	0.065	26^e
	GFLM	81.70	73.67	3.73%	\	2^f; 0.9^g

[a] the number of PLS latent variables.
[b] the penalty parameter in SVM.
[c] the kernel parameter of the Gaussian kernel in SVM.
[d] the number of hidden layer nodes in the MLP model.
[e] the number of hidden layer nodes in the RBF model.
[f] the highest degree of polynomial in the GFLM.
[g] the radii parameter of the subtractive clustering algorithm used in the GFLM.

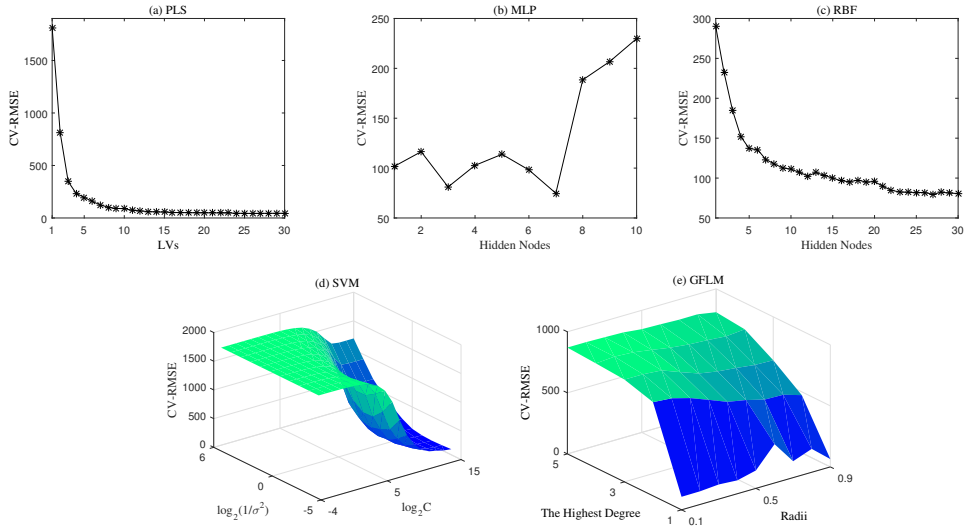

Figure 7: The model hyperparameter cross-validation errors for methane

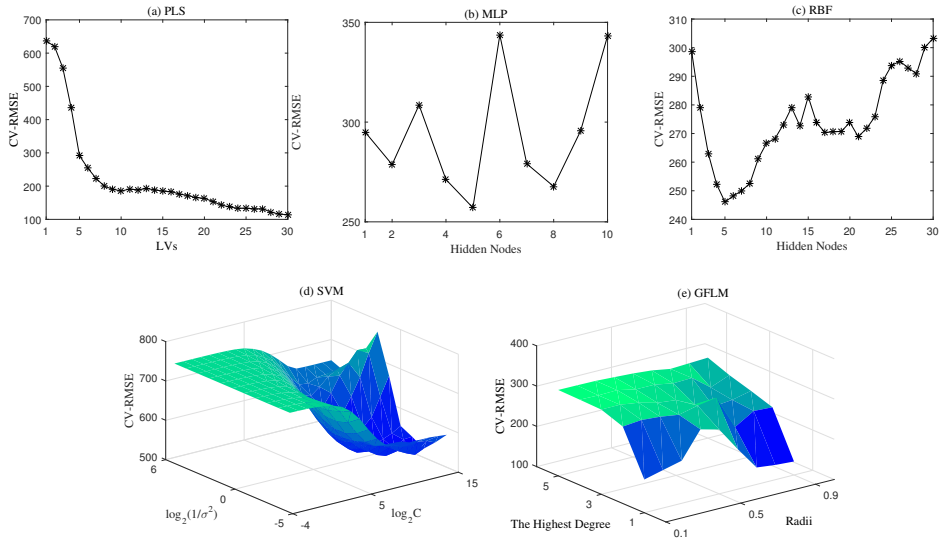

Figure 8: The model hyperparameter cross-validation errors for carbon monoxide

as the learning procedure. The performance of GFLM and PLS, SVM, MLP and RBF models are demonstrated for two real spectral datasets: a coal-fired power plant flue gas dataset and a gas-fired power plant flue gas dataset. The 10-fold cross-validation root mean squares error (CV-RMSE) metric is used to identify optimal values for the hyperparameters of each modelling paradigm. In addition, P-RMSE, REP% and the randomization t-test are employed as metrics to assess the efficiency of the resulting regression models. The results verify that the predictive capability of GFLM is superior to previously reported approaches

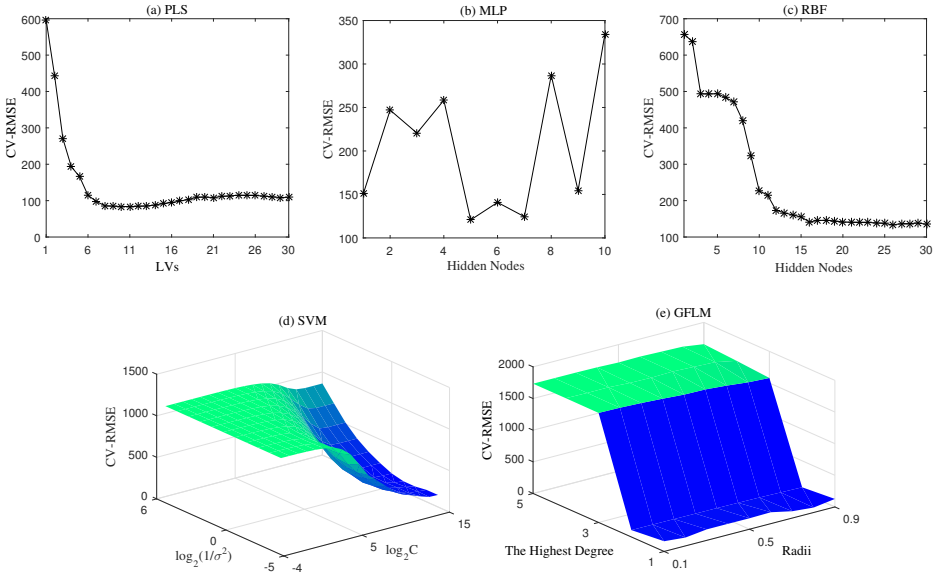

Figure 9: The model hyperparameter cross-validation errors for carbon dioxide

for predicting component concentrations in an optical gas sensing system.

Acknowledgement

This work is supported by National Natural Science Foundation of China (61375055), Program for New Century Excellent Talents in University (NCET-12-0447), Natural Science Foundation of Shanxi Province of China (2014JQ8365), State Key Laboratory of Electrical Insulation and Power Equipment (EIPE16313) and Fundamental Research Funds for the Central University and China Scholarship Council.

References

[1] PRC, Emission Standards for Air Pollutants from Thermal Power Plants GB13223-2011, 2012.

[2] C. T. Driscoll, J. J. Buonocore, J. I. Levy, K. F. Lambert, D. Burtraw, S. B. Reid, H. FakhraeiJ. SchwartzUS power plant carbon standards and clean air and health co-benefitsNat. Clim. Change. 5(2015) 535540.

[3] S. Lakkis, R. Younes, Y. Alayli, M. Sawan, Review of recent trends in gas sensing technologies and their miniaturization potential, Sens. Rev. 34 (1) (2014) 24-35.

[4] S. Sharma, M. Madou, A new approach to gas sensing with nanotechnology, Philos. Trans. R. Soc. A. 370 (2011) 2448-2473.

[5] X. Liu, S. Cheng, H. Liu, S. Hu, D. Zhang, H. Ning, A Survey on Gas Sensing Technology, Sensors, 12 (2012) 9635-9665.

[6] T. Ritari, J. Tuominen, H. Ludvigsen, J. C. Petersen, T. Sorensen, T. P. Hansen, H. R. Simonsen, Gas sensing using air-guiding photonic bandgap bers, Opt. Exp.12 (2004) 4080-4087.

[7] D. Bruno, K. Dmitry, R. Cyril, A. M. Marina, B. Vasily, L. Alexander, L. Andrey, Structure and spectral and luminescent properties of y3al5o12 ceramics containing Ce and Cr, Chemom. Intell. Lab. Syst. 82 (4) (2015) 585-590.

[8] C. Hu, C. Peng, J. Huang, S. Robert, P. Mark, P. Alagappan, A. Daniel, A. Y. Tok, B. Lieberg. Detection of matrilysin activity using polypeptide functionalized reduced graphene oxide field-effect transistor sensor, Anal. Chem. 88(6) (2016) 2994-2998.
[9] F.S. Fedorov, A.S. Varezhnikov, I. Kiselev, V.V. Kolesnichenko, I.N. Burmistrov, Sommer, M. Sommer, D. Fuchs, C. Kuebel, A.V. Gorokhovsky, V.V. Sysoev, Potassium polytitanate gas-sensor study by impedance spectroscopy, Anal. Chim. Acta, 897 (2015) 81-86.
[10] G. Thomas, G, Beate, H. Dale, S.B. Mohammad, M. Makhosazana, S. Aimee, Z. Ralf, A Vacuum Ultraviolet Absorption Array Spectrometer as a Selective Detector for Comprehensive Two-Dimensional Gas Chromatography: Concept and First Results. Anal. Chem. 88(6) (2016) 3031-3039.
[11] C.B. Cai, L. Xu, W. Zhong, Y. Y. Tao, B. Wang, H.W. Yang, M.Q. Wen, Studying a gas-solid multi-component adsorption process with near-infrared process analytical technique: Experimental setup, chemometrics, adsorption kinetics and mechanism. Chemom. Intell. Lab. Syst. 144 (2015) 80-86.
[12] X. Shao, X. Bian, J. Liu, M. Zhang, W. Cai, Multivariate calibration methods in near infrared spectroscopic analysis, Anal. Methods. 2 (2010) 16621666.
[13] D.S. Broomhead, D. Lowe, Multivariable functional interpolation and adaptive networks, Complex Syst. 2 (1988) 321-355.
[14] V. Gaydou, J. Kister, N. Dupuy, Evaluation of multiblock NIR/MI PLS predictive models to detect adulteration of diesel/biodiesel blends byvegetal oil, Chemom. Intell. Lab. Syst. 106 (2011) 190-197.
[15] M. K. D. Rambo, M. M. C. Ferreira, E. P. Amorim, Multi-product calibration models using NIR spectroscopy. 151 (2016)108-114.
[16] P. Geladi1, B.R. Kowalski, Partial least-squares regression: a tutorial, Anal. Chim. Acta. 185 (1986) 117.
[17] D.S. Cao, Y.Z. Liang, Q.S. Xu, Q.N. Hu, L.X. Zhang, G.H. Fu, Exploring nonlinear relationships in chemical data using kernel-based methods, Chemom. Intell. Lab. Syst.107 (2011) 106115.
[18] R. Batruni, A multilayer neural network with piecewise-linear structure and back-propagation learning, IEEE Trans. Neural Netw. 2(1991) 395-403.
[19] D. Zhang, L.F. Deng, K.Y. Cai, Fuzzy nonlinear regression with fuzzified radial basis function network, IEEE Trans. Fuzzy Syst. 13(2005) 742-760.
[20] M. Siamak, F. Tillmann, A.K.S. Johan, Approximate solutions to ordinary differential equations using least squares support vector machines, IEEE Trans. Neural Netw. 23(2012) 1356-1367.
[21] A.K.S.Johan, V.G.Tony, D.B.Jos, D.M.Bart, V.Joos, Least squares support vector machines, World Scientific Publishing, Singapore, 2002.
[22] X.B.Zou, J.W.Zhao, J.W.Malcolm, M.H.Povey,H.P.Mao, Variables selection methods in near-infrared spectroscopy, Anal. Chim. Acta. 667(2010)14-32.
[23] R. Alcala, P. Ducange, F. Herrera, A multi objective evolutionary approach to concurrently learn rule and data bases of linguistic fuzzy-rule-based systems. Fuzzy Systems, IEEE Trans. Fuzzy Syst. 17(2009) 1106-1122.
[24] A.K. Nandi, J.P. Davim, A study of drilling performances with minimum quantity of lubricant using fuzzy logic rules, Mechatronics. 19(2009) 218-232.
[25] M.E. Yuksel, M. Borlu, Accurate segmentation of dermoscopic images by image thresholding based on type-2 fuzzy logic. IEEE Trans. Fuzzy Syst. 17 (2009) 976-982.
[26] F.J. Estrella, M. Espinilla, F. Herrera, A fuzzy linguistic decision tools enhancement suite based on the 2-tuple linguistic model and extensions, Inform. Sciences. 280(2014) 152170.
[27] A. Khosravi, S. Nahavandi, Load forecasting using interval type-2 fuzzy logic systems: optimal type reduction, IEEE Trans. Ind. Inf. 10(2014)1055-1063.
[28] R. Jang, J-S. Sun, C. Tsai, Neuro-fuzzy and soft computing: a computational approach to learning and machine intelligence, Prentice Hall, London, 1997.
[29] Y.X. Wang, H. Cao, Y. Zhou, Y.B. Zhang, Characterization of the flue gas of a coal-fired power plant by ultraviolet-visible spectroscopy and a Tagaki-Sugeno model, Instrum. Sci. Technol.42(2014)576-585.
[30] H. Cao, Y.X. Wang, Nonlinear fuzzy linguistics for the determination of nitrogen monoxide, nitrogen dioxide, and sulfur dioxide by molecular absorption spectroscopy, Instrum. Sci. Technol.43(2015)558572.

[31] K. J. Astrom, B. Wittenmark, Computer Controller Systems: Theory and Design. Prentice-Hall, 1984.
[32] G. C. Goodwin, K. S. Sin, Adaptive filtering prediction and control. Englewood Cliffs, NJ: Prentice-Hall, 1984.
[33] P. Strobach, Linear prediction theory: a mathematical basis for adaptive systems. New York: Springer-Verlag, 1990.
[34] J-S.R.Jang, ANFIS: adaptive-network-based fuzzy inference system, IEEE Trans. Syst. Man, Cybern. Syst. 23 (1993) 665-685.
[35] T. Naes, T. Isaksson, B. Kowalski, Locally weighted regression and scatter correction for near-infrared reflectance data, Anal. Chem. 62 (1990) 664673.
[36] M. Pompe, Prediction of gas-chromatographic retention indices using topological descriptors, J. Chem. Inf. Comput. Sci. 39 (1999) 59-67.
[37] S. Valle, W Li, and S. Joe Qin, Selection of the number of principal components: the variance of the reconstruction error criterion with a comparison to other methods, 38 (1999) 4389-4401.
[38] I.T. Jolliffe, Principal component analysis, Series: springer series in statistics, Springer, New York, 2002.
[39] H Li, C.L. Philip Chen, H Huang, Fuzzy neural intelligent system, Mathematical Foundation and the Application in Engineering, CRC Press, 2000.
[40] H. van der Voet, Comparing the predictive accuracy of models using a simple randomization test, Chemom. Intell. Lab. Syst. 25 (1994) 313-323.
[41] J. Liu, J. M. Hughes-Oliver, J. AM. Jr, Domain-enhanced analysis of microarray data using GO annotations, Bioinformatics. 23 (2007) 1225-34.
[42] Gasmet DX-4000 FTIR gas analyzer on-site series: Instruction and operating manual. 2009.
[43] C. Dumitru, V. Maria. Advantages and disadvantages of using neural networks for predictions. Ovidius University Annals Economic Sciences Series, XIII (2013), 444-449.